

Hybrid MPI/StarSs – a case study

José Gracia*, Christoph Niethammer*, Manuel Hasert[†], Steffen Brinkmann*, Rainer Keller*, Colin W. Glass*

*High Performance Computing Center Stuttgart (HLRS), University of Stuttgart, 70550 Stuttgart, Germany

[†]German Research School for Simulation Sciences GmbH, 52062 Aachen, Germany

Abstract—Hybrid parallel programming models combining distributed and shared memory paradigms are well established in high-performance computing. The classical prototype of hybrid programming in HPC is MPI/OpenMP, but many other combinations are being investigated. Recently, the data-dependency driven, task parallel model for shared memory parallelisation named StarSs has been suggested for usage in combination with MPI. In this paper we apply hybrid MPI/StarSs to a Lattice-Boltzmann code. In particular, we present the hybrid programming model, the benefits we expect, the challenges in porting, and finally a comparison of the performance of MPI/StarSs hybrid, MPI/OpenMP hybrid and the original MPI-only versions of the same code.

I. INTRODUCTION

The Message Passing Interface (MPI) is the *de facto* standard in high performance computing. While MPI is a distributed memory parallelisation scheme, it is also frequently used on SMP-like systems, as for instance today's multi-core CPUs. On such systems, the parallel applications developer may resort to shared memory parallelisation schemes as for instance OpenMP. Hybrid parallelisation models consisting of a distributed memory part to pass messages between compute nodes in a cluster, and a shared memory part to exploit all available cores on a node, have been used in HPC with success. Yet, the general perception is, that MPI-only jobs are still the working horse in research groups relying on numerical computations.

The MPI-only approach has come under pressure; mainly it is challenged by today's prevalence of multi-core and many-core systems. One aspect of the problem certainly is that many applications are structured in such a way that they cannot scale efficiently to large numbers of MPI ranks; they have limits on MPI-scalability that may well depend on problem size or simulation parameters. Often these limits could be lifted or pushed further out by an algorithmic redesign. However, most research groups will *not* go down such a road lightly.

In this paper we will not delve into the lack of scalability of MPI itself, whether perceived or real (see [1] for a discussion on the topic), but will assume that an application's limit on MPI-scalability is inherently a problem of a given application itself. Users hitting the limits of such an MPI-only application have three choices: 1) find a set of parameters (e.g. problem size) that allows to scale out further, 2) redesign and rewrite their application to use MPI more efficiently, 3) go hybrid, which also means redesigning and rewriting the code. Often the first is not possible or interesting from a scientific point of view. From the last two choices, the hybrid approach is usually more straightforward, as the re-design can be implemented

incrementally, allowing to assess its correctness along the way. Note, however, that in view of Amdahl's law, any hybridization needs to cover the code in its completeness in order to be efficient.

The standard paradigm for hybrid parallel programming in HPC is MPI/OpenMP, but many other combinations are being investigated. Recently, the data-flow driven, task parallel model for shared memory parallelisation named *StarSs* has been suggested as a partner for MPI. In this paper, we will first introduce the main aspects of StarSs briefly. Then we will apply the hybrid MPI/StarSs model to an MPI-only Lattice-Boltzmann code while describing the challenges we encountered and how they have been solved. Furthermore, a hybrid MPI/OpenMP version is developed and used as a basis for the assessment of the hybrid MPI/StarSs version. Next, we show performance of the current hybrid implementation, contrast it to the original, and finally draw our conclusions.

II. THE HYBRID MPI/STARSS PROGRAMMING MODEL

We start by giving a short overview of the StarSs programming model [2]. It is based on task parallelisation. First of all, the programmer needs to identify suitable units of work, in general at subroutine boundaries, and designate them as tasks through pragma based code annotations. In contrast to, e.g., pthreads or OpenMP, where task synchronization has to be specified by the programmer explicitly, StarSs infers the synchronization from data dependencies in the program. These dependencies are automatically detected, based on the directionality of subroutine arguments, using code directives to distinguish between input, output and input-output arguments. If programming in Fortran the already present `intent(...)` clauses in subroutine interfaces are used.

At runtime, the data directionalities and the actual data arguments, i.e. their memory address, are taken to build up a dynamic task dependency graph. In the simplest case, tasks will be scheduled for execution sequentially in the order in which they have been added to the graph. This is essentially equal to serial execution. In general, the task graph will expose concurrency which can be exploited by dispatching independent tasks onto a range of available compute cores and running these in parallel. Data dependencies then will ensure that no task is scheduled before all task that modify its inputs have completed. The StarSs model allows to write programs without any kind of explicit synchronisation; in fact, this is the rule. Synchronisation is in principle necessary only when code executed synchronously outside of tasks (executed by a master thread) and code executed inside of tasks (executed

asynchronously by worker threads) use the same data. In this sense synchronisation is required only between sequential code running on the master thread and tasks, not between tasks. Naturally, one can avoid this kind of master-worker synchronization by taskifying all relevant parts of the code.

Compared to classical parallelisation approaches using parallel loops or explicitly synchronised tasks, StarSs has several advantages. Firstly, it is a dynamic process which will allow efficient parallelisation even for different input sets; each run might result in a different dependency graph and thus in different program execution behaviours. Secondly, and much more importantly, the lack of explicit synchronisation between tasks can improve load balancing and solve some of the scalability problems.

The third point to be mentioned concerns the hybrid MPI/StarSs approach; MPI communication can also be off-loaded into StarSs tasks. This allows overlapping of computation and communication in a simple and more efficient way than hand-written code using, e.g., non-blocking MPI operations. In this paper we will explore this possibility only briefly, however.

In some sense, StarSs is a family of programming models including, among others, SMPs and OmpSs. All of them very similar in spirit, particularly the focus on data-dependencies between tasks, but with slight differences in the specifics of the compiler directives and the capabilities of the runtime. For this work, we have used the SMPs compiler, which is tailored to multi-core and SMP-like machines, but also aware of MPI allowing to off-load MPI tasks onto a dedicated communication thread. We did not consider the more recent alternative OmpSs, simply because it did not support Fortran at the time of writing this paper.

III. A CASE STUDY: LATTICE-BOLTZMANN

Before starting to port an application to a new programming model, most application developers in applied computational science will balance expected performance gain against the expected effort. Any new programming model should therefore not only promise high performance, but should also be simple to use. As presented in the previous section, applying the StarSs programming model to an existing application is in principle straightforward as long as the data-flow is clearly exposed in the given code. We have used StarSs to hybridise a previously MPI-only fluid dynamics code which implements the Lattice-Boltzmann Method (LBM). In this paper we present our parallelisation approach, discuss the challenges we encountered, suggest how to overcome them, and finally put the resulting performance into perspective.

We have purposely chosen to undertake this experiment with a code written by a graduate student, which is not heavily optimised regarding scalar or MPI performance. *lbc* [3] – short for *lattice-boltzmann code* – is a Lattice-Boltzmann code using the BGK [4] approximation for the collision term. The algorithm is a standard two-grid approach, i.e. at every algorithmic step the code reads from one grid and writes into the other, without any spatial or temporal blocking. The code

is written in Fortran90 and uses modules, pointers, and user-defined types throughout. Similarly, the MPI parallelisation is based on a simple multi-dimensional domain decomposition. Ghost cells are exchanged at the end of each timestep, no attempt to overlap computation and calculation is done.

The Lattice-Boltzmann method consist of three consecutive computational steps – `stream()`, `collide()`, and `boundaries()`. In order to increase data reuse, the former two are often combined into a `stream_collide()`, while the latter applies external boundary conditions. In the distributed memory case, ghost cells are exchanged with MPI neighbours in `exchange_ghosts()`. In our case the main time-stepping loop is

```
do timestep=1, timesteps
  call stream_collide(block)
  call boundaries(block)
  call exchange_ghosts(comm, block)
end do
```

where `block` is a variable of user-defined type `type(lb)` encapsulating all data structures necessary for the Lattice-Boltzmann algorithm, most importantly the working arrays `flip` and `flop`; `comm` is another user-defined type encapsulating information regarding MPI communication as for instance the rank of neighbours, etc.

A. Tiling of the algorithm with data copies

StarSs is a task parallel programming model. We want to apply it to an algorithm, i.e. LBM, which is as many others in CFD naturally data parallel, but at first glance does not seem to have more than a few obvious subroutines suitable as tasks. Even worse, the candidates `stream()`, `collide()`, `boundaries()`, and `exchange_ghosts()` have flow dependencies on each other and may not be executed concurrently.

In order to apply StarSs to this code, we need to block (or tile) the algorithm, in order to convert data parallelism to task parallelism. In the following we use the term *tiling*. Rather than tiling on the loop level, we do so at the top program level by introducing an additional domain decomposition for tiling on top of the MPI domain decomposition. In the following we will refer to an MPI subdomain as *block* and to a StarSs subdomain as *tile*. Note, that tiles are subdomains of a particular block. In the code, the variable `block` represents a block and tiles are represented by `tile`. For simplicity each tile is based on the same user type as blocks allowing to reuse much of the original code.

Similar to blocks on the MPI level, tiles need to exchange ghost cells with every neighbour. So, before doing the LBM step, tiles pull ghost cells from a buffer (`pull_ghosts()`) and push them back into a buffer (`push_ghosts()`) after concluding the LBM step. We have chosen to use the MPI level block as buffer. The tiled main time-stepping loop is then:

```
call init_tiles(block, tiles)
```

```

do timestep=1, timesteps
  do iTile=1, nTiles
    tile => tiles(iTile)
    call pull_ghosts(tile , block)
  end do

  do iTile=1, nTiles
    tile => tiles(iTile)
    call stream_collide(tile)
    call push_ghosts(tile , block)
  end do

  call boundaries(block)
  call exchange_ghosts(comm, block)
end do

call gather_tiles(block , tiles)

```

Note, that inside the time-stepping loop `block` only carries valid data for those array elements that are required to exchange ghost cells, either with MPI neighbours in `exchange_ghosts()` or with neighbour tiles in `push/pull_ghosts()`, respectively. At the end of the timestep loop, or before any output for that matter, data in the tiles needs to be gathered back into the block, which is done in `gather_tiles()`. Similarly, `init_tiles()` initialises tiles from data in the block.

B. Tiling of the algorithm with data aliasing

A major drawback of the tiled algorithm described in the section above, is the data copies between the tiles and the block. This is true even if only those array elements are copied which serve as ghost cells for adjacent tiles or for neighbouring MPI blocks. The resulting code suffers significant performance penalties in the order of several tens of percent, as will be detailed later on.

Rather than replicating block's data in the tiles, we sought a way of reusing the data in a block, by mapping or aliasing it into the tiles. We made use of the fact that in Fortran, array pointers can not only point to whole arrays, but also into sub-arrays as

```

tile%flip => block%flip(.. ,xl:xu, ..)
tile%flop => block%flop(.. ,xl:xu, ..)

```

where the array boundaries in the block, i.e. `xl`, `xu`, etc, are distinct for each tile. Note, that in general the pointers will alias sub-arrays which are non-contiguous in memory. With this implementation tiles and the block are just different views on the same underlying memory region. The data is organised in such a way, that it simplifies data access in different parts of the code, respectively.

Obviously, it is no longer necessary to keep tiles and blocks consistent, nor to copy ghost cells between adjacent tiles. The subroutines `pull_ghosts()`, `push_ghosts()`, and `gather_tiles()` are no longer needed. The only purpose of `init_tiles()` is to setup the aliases for `flip` and `flop`. The basic algorithm thus simplifies considerably to

```

call init_tiles(block , tiles)

do timestep=1, timesteps
  do iTile=1, nTiles
    tile => tiles(iTile)
    call stream_collide(tile , block)
  end do

  call boundaries(block)
  call exchange_ghosts(comm, block)
end do

```

Note, that this code resembles very closely the original non-tiled version. The only difference apart from the initialisation of tiles is the additional tiles loop embedding the call to the main computational step `stream_collide()`. The inclusion of the argument `block` to task `stream_collide()` will become clear in section III-D.

C. Task identification and data directionality

All subroutines in the main time-stepping loop are in principle suitable as task. In StarSs neither scheduling of tasks nor synchronisation amongst them is specified explicitly. Instead, the developer has to declare the directionality for each argument of any task subroutine, i.e. input, output, or inout. In C, this is done through additional clauses to the StarSs task directive. The Fortran interface uses the Fortran clause `intent(...)`, which becomes mandatory for each argument of task subroutines. The interface of our tasks is:

```

interface
  !$CSS TASK REDUCTION(sentinel)
  subroutine stream_collide(tile , sentinel)
    type(lb), intent(inout) :: tile
    type(lb), intent(inout) :: sentinel
  end subroutine
  !$CSS TASK
  subroutine pull_ghosts(tile , block)
    type(lb), intent(inout) :: tile
    type(lb), intent(in)    :: block
  end subroutine
  !$CSS TASK REDUCTION(block)
  subroutine push_ghosts(tile , block)
    type(lb), intent(in)    :: tile
    type(lb), intent(inout) :: block
  end subroutine
  !$CSS TASK
  subroutine boundaries(block)
    type(lb), intent(inout) :: block
  end subroutine
  !$CSS TASK
  subroutine exchange_ghosts(block)
    type(lb), intent(inout) :: block
  end subroutine
end interface

```

Note, that without the reduction clause, i.e. `REDUCTION(block)`, on task `push_ghosts()` the

application would be partly serialised: each invocation of `push_ghosts()` would depend on the previous one through the argument `block` with `intent(inout)`. Thus, each instance of this task function could not be executed before the previous one completed execution. The reduction clause instructs the StarSs runtime to ignore `inout`-dependencies on a given datum for tasks participating in the reduction, but retains dependencies to tasks not being part of the reduction. It is then the responsibility of the developer to make sure that parallel execution of reduction tasks does not lead to data-races, e.g. by making sure that different tasks invocations modify only non-overlapping data regions, or by using locks or atomic instructions.

It is worth noting that – conceptually – the most difficult part of using StarSs is constructing data dependencies in a way, that allows sufficient concurrency to be identified and therefore exploited by the runtime. Another difficulty is verifying that the directionalities specified by the developer through StarSs directives are compatible with the actual behaviour of the code. In both cases the StarSs debugger TEMANEJO [5], [6] is a very useful tool. It shows a visual representation of the actual dependency graph at runtime and has the ability to control the scheduling of tasks.

D. Fortran issues

This study was done using the SMPSSs flavour of the StarSs programming family, more specifically SMPSSs V2.5 compiler and runtime [7] were used. Unfortunately, this version still suffers from a couple of problems regarding Fortran90 support.

The most important issue is that subroutines declared in modules may not be used as task functions, i.e. task functions are not allowed to be part of a module. We have worked around this issue, by declaring a wrapper subroutine outside of any module. The wrapper basically just calls the original subroutine after `USEing` the respective module.

The next issue is that the SMPSSs compiler and runtime system are not able to track dependencies on a subset of data, in the sense of a sub-array `a(2:3)` being a subset of the full array `a(:)`. Currently, SMPSSs has no notion of data *size* (at least when it comes to dependency-tracking). If one task has `intent(out)` on `a` and another one has `intent(in)` on `a(2:3)`, the runtime will not acknowledge a dependency between those two tasks, as it only conceptualizes data based on its starting address in memory. A similar situation arises in our case: each StarSs tile is a subset of the MPI block (even if there is no overlap in memory of the variables `tile` and `block`), but there is no way to convey this information to the runtime.

In our case, the Fortran modules were no real problem. Using wrappers as workaround was simple. A shortcoming resulting from this layer of wrappers across Fortran modules is, that it prevents the compilers from doing certain optimisations. The performance penalty in our case was at the sub-percent level and thus acceptable.

It is difficult to give a general solution to the data subset problem. One approach that has been suggested [8] is to use

sentinels, sometimes also called representants. One datum, or rather its memory address, is chosen to represent a whole collection of data. This is similar to colouring array elements in loops. The sentinel is then passed as additional argument to tasks subroutines with a corresponding intent. In principle the sentinel can be any variable, in the simplest case an array with a single element. Sometimes it is preferable to use an existing variable in order to prevent clobbering the source code with additional variables without apparent function. For instance, we have used `block` as a sentinel and passed it to tasks without actually using it inside the task subroutines. This is the only reason, why our task `stream_collide()` takes `block` as argument. Note, that a reduction on `block` is necessary to allow all instances of this task to run in parallel; omitting this would result in a serialization of all tasks. Note also, that in the version presented in section III-A no sentinels are required as the task `push_ghost()` connects all tiles with the block dependency-wise.

E. Performance measurements

We have benchmarked three different versions of the applications. All versions were compiled with GNU gfortran 4.6.1 as (backend) compiler. The first is the original MPI-only version without any modifications for StarSs and was compiled with the Open MPI 1.5.4 compiler driver. The second, *MPI/StarSs or hybrid version*, is the version described in this paper, i.e. has been refactored to use tiles with aliasing (as describes in section III-B) and works around the Fortran issues above. Throughout this paper we use a fixed tilesize of 32^3 lattice nodes. This version has been compiled with the SMPSSs 2.5 compiler driver. Initial performance measurements of the tiled version using data copies (see section III-A) showed a penalty of roughly 40% with respect to the aliased version (depending on tile- and total problem size) and was not considered further. The third version is a hybrid *MPI/OpenMP* version. It uses the same tiled code base as the StarSs version, but does a `omp parallel do` on the tiles loop, which implies a synchronisation before applying boundary conditions.

The benchmarks have been performed at HLRS on the cluster *Laki*[9]. *Laki* is a cluster of roughly 700 nodes interconnected through an Infiniband network. Each node is a two-way NUMA node and has two Nehalem X5560 sockets, with 4 cores each, resulting in 8 cores per node. Each node has access to 12GB of memory, however, non-uniformly across sockets.

The MPI ranks were placed in such a way that communication to off-node destinations was the same for all three versions. Communication with on-node partners was done explicitly through library calls for the MPI version. In the OpenMP version, data is shared and synchronization done through explicit OpenMP directives, while the StarSs version is self-synchronised through implicit data dependencies.

Single-node strong scaling

The first experiment we carried out was to measure the behaviour of the codes under strong scaling on a single node.

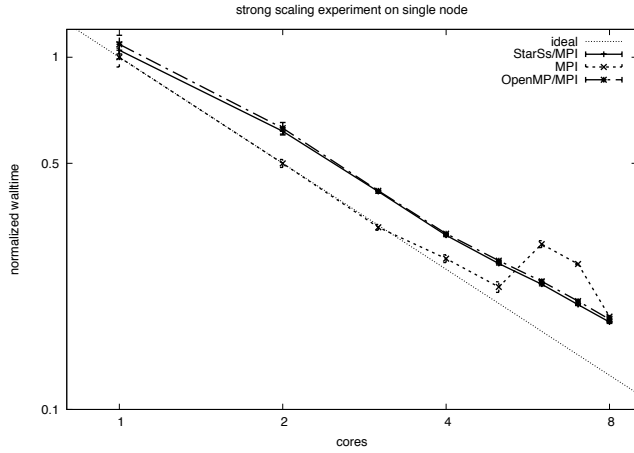


Figure 1. Strong scaling experiment on a single node. Values are normalized to the execution time of the MPI-only version on a single core.

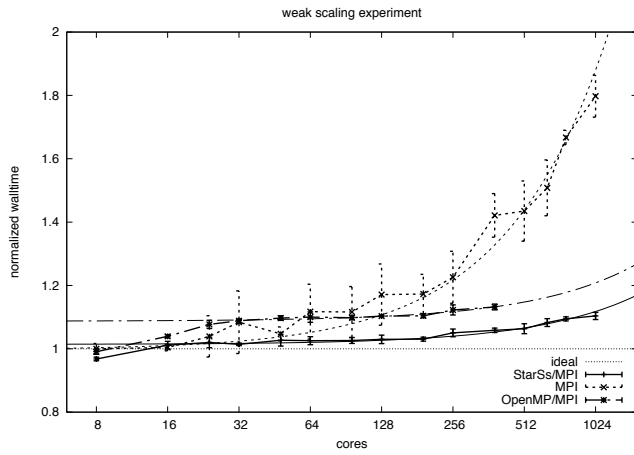


Figure 2. Weak scaling experiment on a cluster of Nehalem nodes plotted over number of cores. Values are normalized to the execution time of the MPI-only version on 8 cores.

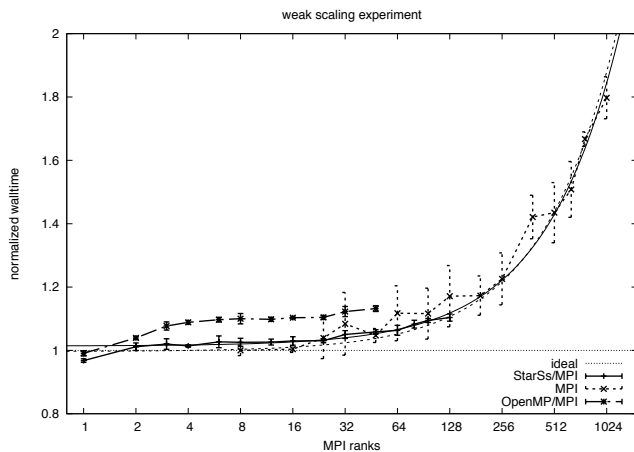


Figure 3. Weak scaling experiment on a cluster of Nehalem nodes plotted over number of MPI ranks (as opposed to cores in Fig. 2). The same normalization as in Fig. 2 is used.

The result for a $256 \times 256 \times 224$ lattice is shown in Fig. 1. This is the largest problem size that can be run on a single Laki node. We have timed 50 time-steps, including the initialisation phase and a single phase of output at the very end. Each measurement point is the mean of a large number of runs; runs were done in several batches on different days in order to sample a range of realistic load patterns of the cluster. The measurement error is taken to be the statistical standard deviation. We have normalised all measurements to the execution time, $t_1 = 115.04s$, of the MPI-only version running on a single MPI rank. In the Lattice-Boltzmann community it is good practice to compare the performance of different codes by the number of lattice-node updates per second (MLUPs, mega lattice updates per second); our normalisation then corresponds to 6.4 MLUPs per core¹.

The first thing to notice from Fig. 1 is that on a single core there is no significant difference in the performance of the three version, as the difference is smaller than the measurement errors. Looking past measurement errors, one can see a performance drop from MPI through MPI/StarSs to MPI/OpenMP of less than 8%.

Doubling the number of cores from one to two, the MPI version scales nearly perfectly, while both hybrid versions have significantly lower scaling efficiency. Both hybrid versions, i.e. MPI/StarSs and MPI/OpenMP, scale with high efficiency beyond two cores up to the full complement of 8 cores, but remain below the performance of the MPI version up to 6 cores. The efficiency of the MPI versions initially is very good, but then drops noticeably beyond 4 cores. In fact, there is only little overall speedup from 5 to 8 cores with an intermediate increase in execution time. When 8 cores are reached, both hybrid versions have made up on the efficiency they lost initially compare to the MPI version and perform slightly better. This slight performance advantage is not significant for the MPI/OpenMP version. The performance advantage of the MPI/StarSs version over MPI-only is slight, but statistically significant.

Of the three versions, the MPI-only version displays the largest scatter in execution times by a factor of at least three compared to the hybrid versions. The measurement error of the hybrid StarSs version is relatively high for a single and two cores, but then falls off noticeably.

The hybrid versions do not seem to be greatly affected by the NUMA nature of the *Laki* nodes, nor the limited memory bandwidth, at least not more than the MPI version. In fact, pinning MPI processes to cores or sockets does not increase performance for the MPI version and is difficult² to implement particularly for the hybrid versions. We thus decided to use 8 threads per MPI process for the hybrid versions.

¹These values are lower than those quoted in the literature for highly optimised LBM codes. Gains in scalar performance have the price of higher complexity of the source code which we wanted to avoid in this paper. Yet, all techniques discussed here can in principle be applied to highly optimized codes as well.

²in the sense of not easily portable across platforms

Multi-node weak scaling

A weak scaling experiment with $32 \times 256 \times 224$ lattice nodes per core was done; core counts range from 8 (a single node) to 1024 cores. Again we have normalised all measurements to the execution time of the MPI-only version, this time running with 8 cores, i.e. $t_8 = 21.13s$, which corresponds to 4.3 MLUS per core.

The results of the weak scaling experiment are presented in Figs. 2 and 3. Also in this batch of runs, the hybrid MPI/StarSs version tends to be slightly faster than the MPI version, albeit only with a low significance. Increasing the number of cores, the execution times of all three versions can be modelled by smooth curves as discussed in the next section IV. No sudden change of scaling efficiency is observed in any of the versions. Only the MPI/OpenMP version scales at lower efficiency below 32 cores, incurring a performance penalty of 10% compared to the MPI/StarSs version. The scaling efficiency of the hybrid MPI/StarSs version is significantly better than the one of the MPI-only version. At 1024 cores the execution time of the MPI version has gone up by a factor of roughly 1.8 while the hybrid version is still running within a factor 1.1 of the ideal execution time. So far, we have not been able to collect sufficient measurements for the MPI/OpenMP version beyond 384 cores. However, the data we have seems to indicate that also this version scales considerably better than the MPI-only version, but due to the low scaling efficiency at small numbers of cores, it ends up running at a factor of only 1.2 of the ideal execution time.

As with the strong-scaling experiment, the scatter of execution times is significantly larger for the MPI than for the hybrid versions; typically by a factor of 5 or higher.

IV. DISCUSSION

A. Multi-node scaling efficiency

In this section we will discuss the experimental results presented in the previous section in some more detail, particular in terms of scaling efficiency. To quantify this, we start out by modelling the total execution time $T(n)$ of the weak-scaling experiments by a parametric function in the number of cores n given as

$$\frac{T^i(n)}{T_8} = \tau_0^i + \mu^i n, \quad (1)$$

where the index i denotes the three versions (m, s, o for MPI, MPI/StarSs, and MPI/OpenMP, respectively); T_8 is the normalization of the weak-scaling experiments (chosen to be the execution time of the MPI-only version at 8 cores). The parameter τ_0^i quantifies a constant contribution to the execution time, and μ^i one that scales linearly in the number of cores. Fits of the measurements to this parametric function are shown together with the data in Fig. 2. Higher order functions with terms proportional to $n \log n$, n^2 , etc, did not improve any of the fits significantly.

From the fit, the linear contribution to the execution time of the MPI/StarSs version is $\mu^s = (1.0 \pm 0.1) \times 10^{-4}$, for the MPI-only version it is $\mu^m = (8.6 \pm 0.2) \times 10^{-4}$, for the

MPI/OpenMP version it is $\mu^o = (1.2 \pm 0.2) \times 10^{-4}$; values are rounded to the nearest digit. This means, that the execution time of the weak-scaling experiment will have doubled with respect to the reference (i.e. 8 cores) at 1160 cores for the MPI version, at 8470 cores for the hybrid MPI/OpenMP version, and at 9860 cores for MPI/StarSs.

The ratio is $\mu^m/\mu^s = 8.5 \pm 0.5$, which is the number of StarSs threads running inside an MPI process for the hybrid version. In other words, the hybrid version scales, in terms of MPI processes, exactly the same way as the MPI-only version. However, it uses all of the potential of the hybrid approach and scales a factor of 8 more efficiently in terms of number of cores. This circumstance can be nicely illustrated by plotting the performance data over the number of MPI ranks (see Fig. 3) rather than over number of cores (see Fig. 2). In terms of MPI ranks, the MPI-only and the hybrid MPI/StarSs version scale nearly identically. Compared to MPI/StarSs, the MPI/OpenMP version scales slightly less efficient with $\mu^m/\mu^o = 7.3 \pm 1.0$ yet not significantly.

Under weak-scaling, the time a single core spends doing actual computation is constant, i.e. independent of the total number of cores. The total execution time may deviate from this ideal value due to communication time. Our parametric function eq (1) can thus be interpreted in terms of a constant computation time τ_0^i (including any constant-time communication contribution), plus a communication term $\mu^i n$ which is proportional to the number of cores. As the MPI domain decomposition is one-dimensional, the linear communication time found in our scaling curves suggest, that in the current MPI version of the code, the next-neighbour communication is serialized across the communicator. And indeed, the original code uses the same rank as source and destination argument for the `MPI_sendrecv()` call. Obviously, this is simple to remedy and would make the communication time a constant, independent of the number of cores. Note, that this statement is true for all three versions of the code as it affects only the communication between MPI ranks, and does therefore not change the relative performance of the different versions.

B. Single-node performance

Next we would like to address the single-node performance of the different code versions. As noted above, the performance of the three versions on a single node is very similar. Assuming that measurement errors were negligible, the StarSs version is roughly 5% faster than the MPI version. These two versions differ in 1) the runtime environment, and 2) the tiling in the code. Any of these two could be causing the difference in performance.

We have therefore compiled the tiled code (as used for the StarSs version) with the same MPI wrapper as our original MPI-only version. Given the measurement errors, the single-core performance of this tiled MPI version is indistinguishable from the single-core performance of the other three versions discussed in this paper. At this level we cannot measure any statistically significant effects of the respective runtimes, i.e. any overheads of SMPSSs are negligible (within the established

measurement error estimates) for the problem size and task granularity under consideration.

At 8 cores, however, the hybrid MPI/StarSs and the tiled MPI version have a significant performance advantage over the original non-tiled MPI code. While our data hints at a slight advantage of the tiled MPI version also over the tiled hybrid version, the difference is not significant.

We have modelled the total execution time $T(n)$ of the strong scaling experiment by a parametric function in the number of cores n as

$$\frac{T^i(n)}{T_1} = \frac{\tau_0^i}{n} + \beta^i + \eta^i (n - 1), \quad (2)$$

where the index i denotes any of the three versions (m, s, o for MPI, MPI/StarSs, and MPI/OpenMP, respectively); T_1 is the normalization of the weak-scaling experiments (chosen to be the execution time of the MPI-only version at 1 core). The parameters τ_0^i relates the execution time of a particular version to the reference, β^i parametrizes constant-time contributions as for instance memory bandwidth requirements that cannot be parallelized, and η^i quantifies deviations from the ideal speedup that scales linearly in the number of cores. This parametric function was used to fit the measurements of the strong scaling experiment. Terms proportional to $n \log n$, n^2 , etc, did not improve any of the fits significantly.³ In all cases the parameter β^i is not well constrained by the data and consistent with $\beta^i = 0$. We interpret the term $\eta^i (n - 1)$ as a measure for contention amongst the cores during accesses to the memory system, either the cache or the main memory. The ratios $\eta^m/\eta^s = 2.3 \pm 0.2$ and $\eta^m/\eta^o = 1.8 \pm 0.2$, indicate that the hybrid versions scale significantly better on a single node than the MPI-only version. If our interpretation of the parameter η^i is correct, we conclude that the MPI version suffers stronger from memory contention.

Our application is memory bandwidth limited – a resource that decreases in proportion to the number of cores that need to share it. The higher re-use of cached memory lines by the tiled algorithm leads to a more efficient use of memory bandwidth. We thus argue that high-level tiling of the algorithm is beneficial from a cache re-utilization perspective⁴ also for the MPI-only version and not just an up-front investment with the sole purpose of applying the task-based programming model StarSs. In this view, the porting effort of StarSs, i.e. code annotations for task identification and directionality of arguments, is minimal.

Finally, we would like to address the peculiar performance measurements of the MPI-only version between 5 and 7 cores. The drastic increase in execution time stems from an imbalanced domain decomposition; the domain (a power of 2) simply cannot be divided into equal sub-domains for 5, 6, or 7 cores. The same is true for 3 cores also, but there the load-imbalance is relatively small.

³Note, that for fitting the MPI version we have only taken into account core numbers equal to powers of two.

⁴Naturally, this high-level tiling of the algorithm may not substitute low-level loop-blocking tailored to specific cache hierarchies.

C. Overlapping computation and communication

One of the goals when optimizing MPI code is to hide the communication latency by overlapping computation with communication. In our application MPI calls can be issued as soon as the outermost lattice nodes of the domain have been calculated. One possible implementation is to split the low-level loops in the computation subroutines to do the necessary iterations first, issue non-blocking MPI calls, and only then do inner loops. In general this will lead to obfuscated, bloated code, in particular if there are many computation subroutines.

An algorithm which is tiled on a high semantic level – as ours – suggests a different approach. Since all calculations on a specific `tile` are done with a single subroutine call, it is sufficient to reorder a single loop, namely the one over all tiles. One still has to use non-blocking MPI and wait on the messages at a suitable point in the code.

In our hybrid MPI/StarSs version we have followed a variation of this scheme. We introduced two disjoint sentinels, representing the outermost tiles and the remaining inner tiles respectively, and called the computation routine `stream_collide()` with the proper sentinel. The subroutines `boundaries()` and `exchange_ghosts()` then carry only the sentinel representing the outer tiles. In this way we arrive at a dependency tree, that allows to issue boundary conditions and MPI communication as soon as computations on the outer tiles are done; inner tiles are calculated in parallel. No modifications to the original blocking MPI calls are necessary.

The amount of communication time that can be hidden depends on the ratio of inner to outer tiles and the number of cores working on them. Specifically, we can hide roughly 250 ms communication time per timestep. As the communication time increases with the number of cores, this limit is reached between 128 and 256 cores. Below that, communication is completely overlapped with calculations. With this communication scheme, fitting the parametric function eq. 1 to data points between 32 and 256 cores, results in a much flatter curve for the MPI/StarSs version; with $\mu^s \sim 3 \times 10^{-5}$ the total execution time would double only at beyond 32000 cores based on an extrapolation of the curve. This makes the hybrid MPI/StarSs version scale much better than the MPI-only version in this range.

V. CONCLUSIONS

We set out to investigate if the hybrid MPI/StarSs programming model allows typical applications to scale to larger number of cores than MPI-only applications and how much porting effort this model requires. Most research groups applying computational methods to solve their domain specific problems will have only limited resources to do such a port.

In our LBM example we have shown that the hybrid version, combining MPI with the task-based programming model StarSs, is a) competitive with the MPI-only version on a single node, in fact outperforms it, b) leverages the full scaling efficiency across nodes, and c) results in much more stable execution times reducing the effect of load-imbalance

(internal or external as, e.g., due to OS jitter). We have also evaluated a hybrid MPI/OpenMP version based on the same code refactoring done for MPI/StarSs.

Both hybrid versions are competitive with the MPI-only version on a single core regarding strong scaling. The overhead of StarSs task instantiation, management, and scheduling is negligible for the chosen task granularity. In fact, the StarSs version is slightly faster than the MPI version when the full node is used. This stems from a better cache re-use due to tiling which leads to a lower contention among cores on the memory system. The same is true for the hybrid MPI/OpenMP version but to a somewhat lower degree.

Doing hybrid distributed/shared memory parallelisation is often motivated by a reduction of MPI ranks involved in a calculation in order to reduce communication times which depend on the number of ranks involved. At the very least, hybrid models should allow to scale a given application to a larger number of cores than the MPI-only version. Ideally, the range to which an application scales is extended by a factor equal to the number of threads running inside a single MPI rank. Scalar overheads of the hybrid model should remain small in order not to counteract the higher scaling efficiency – doing so increases the parallel efficiency of the application to a much larger number of total *cores*, in particular on multi- or many-core systems.

We have shown that the hybrid versions, in particular the MPI/StarSs, make up on their promise and leverage the full potential by allowing to scale out by a factor equal to the number of cores on a single node. In addition, StarSs allows to hide communication latencies in a straightforward way. In our case this turned out not to be very relevant, as the communication time increases very rapidly and eventually can not be hidden. Up to this stage, however, the weak scaling curve is practically flat. At scale, it is more efficient to use the hybrid version than the MPI-only version.

In a future paper we will investigate different methods to overlap communication and computation in more detail. In StarSs this is particularly attractive as the dynamic scheduling driven by data dependencies allows to hide communication latencies as long as sufficient concurrency is available. Also, the SMPSS runtime specifically allows to off-load MPI communication onto a special communication thread [10] that, in most cases, will not use cycles and allows to have multiple transfers on the fly at the same time. In that case the programmer does not need to bother with non-blocking asynchronous MPI calls, but may continue to use the most simple MPI calls and still get asynchronous communication with latency hiding.

The MPI version has a much larger scatter of execution times for the weak and the strong scaling experiments than both of the hybrid versions. While we do not have execution traces to support our hypothesis, we argue that the scatter stems from load imbalances due to OS jitter. All code versions compete for resources with the OS, however, any delay caused to a single task of the hybrid versions will not affect the tasks running on the remaining cores, yet for the MPI version all cores have to idle eventually waiting on the delayed core to

finish its work thus multiplying the initially small disturbance. In any case hybrid models help with application specific load imbalances.

Finally, the porting effort was limited to tiling the algorithm, which is not particularly difficult for LBM codes, especially in Fortran where aliasing of sub-array can be used. Given the fact, that the MPI-only version also benefits from this tiling, it is very much worth the effort. Once done, executing the code in hybrid mode, either in combination with OpenMP or StarSs allows to scale out to much larger number of cores. A future work will investigate if and how StarSs can be used to do temporal blocking in addition to the spatial blocking presented in this work.

VI. ACKNOWLEDGMENTS.

This work was supported by the European Community's Seventh Framework Programme [FP7-INFRASTRUCTURES-2010-2] project TEXT under grant agreement number 261580 and by the project PRACE-IIP under contract RI-261557.

REFERENCES

- [1] P. Balaji, D. Buntinas, D. Goodell, W. Gropp, S. Kumar, E. Lusk, R. Thakur, and J. L. Träff, "Mpi on a million processors," in *Proceedings of the 16th European PVM/MPI Users' Group Meeting on Recent Advances in Parallel Virtual Machine and Message Passing Interface*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 20–30.
- [2] P. Bellens, J. M. Perez, R. M. Badia, and J. Labarta, "Cellss: a programming model for the cell be architecture," in *ACM/IEEE CONFERENCE ON SUPERCOMPUTING*. ACM, 2006, p. 86.
- [3] M. Hasert, H. Klimach, and S. Roller, "Caf versus mpi - applicability of coarray fortran to a flow solver," in *Lect. Notes. Comput. Sci.* Springer, 2011.
- [4] P. L. Bhatnagar, E. P. Gross, and M. Krook, "A Model for Collision Processes in Gases. I. Small Amplitude Processes in Charged and Neutral One-Component Systems," *Phys. Rev.*, vol. 94, no. 3, pp. 511–525, 1954.
- [5] S. Brinkmann, C. Niethammer, J. Gracia, and R. Keller, "Temanejo - a debugger for task based parallel programming models," in *ParCo2011: Proceeding of the International Conference on Parallel Computing*, 2011, accepted.
- [6] "TEMANEJO - a starss debugger," Jul. 2011. [Online]. Available: <http://www.hlrs.de/organization/av/amt/projects/text/>
- [7] (2011, Jul.) Smpss distribution v2.5. [Online]. Available: <http://www.project-text.eu/sites/default/files/SMPSS-2.5.tar.gz>
- [8] J. Perez, R. Badia, and J. Labarta, "A dependency-aware task-based programming environment for multi-core architectures," in *Cluster Computing, 2008 IEEE International Conference on*, 29 2008-oct. 1 2008, pp. 142–151.
- [9] "Nec nehalem cluster," Jul. 2011. [Online]. Available: <http://www.hlrs.de/systems/platforms/nec-nehalem-cluster/>
- [10] V. Marjanović, J. Labarta, E. Ayguadé, and M. Valero, "Overlapping communication and computation by using a hybrid mpi/smpss approach," in *Proceedings of the 24th ACM International Conference on Supercomputing*, ser. ICS '10. New York, NY, USA: ACM, 2010, pp. 5–16. [Online]. Available: <http://doi.acm.org/10.1145/1810085.1810091>